

1 The role of cross-linguistic lexical similarity on bilingual word acquisition

2 Gonzalo Garcia-Castro¹, Daniela Avila-Varela¹, & Nuria Sebastian-Galles¹

3 ¹ Center for Brain and Cognition, Universitat Pompeu Fabra, Barcelona, Spain

4 Author Note

5 Campus de Ciutadella, Universitat Pompeu Fabra, 08005, Barcelona, Spain

6 Correspondence concerning this article should be addressed to Gonzalo

7 Garcia-Castro, Ramon Trias Fargas, 25-27, 08005 Barcelona, Spain. E-mail:

8 gonzalo.garciadecastro@upf.edu

Abstract

Bilinguals face the challenging task of learning words from languages with overlapping phonologies. Floccia et al. (2018) reported larger vocabulary sizes for 24-month-old bilinguals that were learning languages that shared a greater amount of cognates (e.g., English-Dutch). The mechanisms underlying this effect remain unknown. We explore two compatible scenarios. First, we test whether cognates are learnt earlier than non-cognates. This would account for the difference in vocabulary size associated to the amount of shared cognates across languages. Second, we explore the possibility that the word-forms of one language interact with those from the other language, scaffolding the acquisition of their translation equivalents when their phonologies overlap. This mechanism, in line with the parallel activation account of bilingual speech perception, would provide a plausible explanation to why cognates are acquired earlier by bilinguals. We developed an online tool to collect parental reports of receptive and productive vocabularies from children learning Catalan and/or Spanish, and present data on receptive and productive vocabulary of bilingual toddlers aged 12 to 34 months.

Keywords: lexical acquisition, vocabulary, bilingualism

Word count: X

The role of cross-linguistic lexical similarity on bilingual word acquisition

Introduction

Learning words involves two key steps: (1) the encoding of a word form (e.g, /d /) and (2) its association with a referent (e.g., a DOG). There is evidence that word learning occurs at early ages: six-month-old infants show a preference toward named pictures, relative to unnamed pictures displayed side-by-side (???). At 7.5 months, the phonological representations of familiar words seem to be quite detailed, as infants show evidence of recognition when the whole word is uttered (???), as opposed to when only part of the word is uttered. At 12 months, infants show sensitivity to both vowel and consonant mispronunciations (???), looking shorter to named pictures when their label is mispronounced. Remarkable as it is, the phonological specificity of infants' lexical entries is only the foundation of following developmental milestones. One of them, central to this study, is the emergence of excitatory and inhibitory links between word representations, which represents one of the essential characteristics of the adult lexicon (???).

Words are not learnt in isolation, but in contexts rich in word tokens and referents. The structure of toddlers' lexicon reflects this high connectivity (???; ???; ???). At 18 months, infants' recognition of spoken words is sensitive to phonological priming effects, suggesting that their lexical entries are phonological linked, and co-activated during speech processing (???; ???). The emergence of semantic links in the developing lexicon occurs somewhat later, at 24 months of age, when infants show sensitivity to semantic priming (???; ???; ???), and show evidence of inhibitory links between lexical-semantic entries (???). By 24 months, toddlers' recognition of familiar words also seems to follow a hierarchical fashion, as revealed by their sensitivity to the phonology and semantic similarity of distractor pictures presented along the named picture: the interference of phonological distractors is stronger at earlier stages of spoken word recognition, that of semantic distractors is stronger at later stages (???). Vocabulary size predicts the

emergence and the strength of these effects: 18 months-old toddlers are sensitive to semantic priming and interference too, provided they know a similar amount of words (???, ???), and the strength of the phonological and semantic distractors during spoken word recognition grows stronger along the vocabulary size of the toddlers. This suggests that the emergence of rich lexical networks – both at the phonological and the semantic levels – grows as a function of the number of words acquired by the toddler.

The structure of the lexicon also impacts the order of acquisition of new words. Words that share a high degree of phonological and/or semantic overlap with words already acquired are more likely to be acquired next (Hills, Maouene, Maouene, Sheya, & Smith, 2009). (???) and (???) showed that the connectivity of a given semantic category (i.e., animals) in a child’s lexicon predicts a better performance in a disambiguation task, where participants are presented with a novel label in the presence of a familiar and a novel object: infants show stronger looking preference for the novel object if it belonged to a category for which many words had already been acquired. This points to the structure of the lexicon facilitating the strategies children engage during word learning.

The case of bilingual children (here defined as those learning two languages from birth) presents an opportunity to study how lexical acquisition takes place under a more complex environment, as they face the challenge of learning two distinct sets of words – one for each language – that partially overlap in sound and meaning.

EARLY STAGES IN BILINGUAL LANGUAGE ACQUISITION

When compared to their monolingual peers, they know fewer words only when just one language is considered (e.g., English monolinguals know more words in English than English-Spanish bilinguals). When both languages are taken into account, bilingual children seem to know, at least, as many words as monolinguals do (???, ???; Ben-Zeev, 1977; Bialystok, Luk, Peets, & Yang, 2010; Blom et al., 2019; Core, Hoff, Rumiche, & Señor, 2013; Fernandez, Pearson, Umbel, Oiler, & Molinet-Molina, 1992; Hoff et al., 2012;

Hoff, Rumiche, Burridge, Ribot, & Welsh, 2014; Pearson & Fernández, 1994; but see Pearson, Fernández, & Oller, 1993; Houwer, Bornstein, & Putnick, 2014).

Floccia et al. (2018), gathered parental vocabulary estimates (Hamilton, Plunkett, & Schafer, 2000) from a large sample of 24-month-old toddlers learning English and an additional language (from a pool of 13 linguistically diverse languages). They computed the average phonological similarity between translation equivalents (TEs from now on; e.g., *table* in English and *tafel* in Dutch) for each pair of languages, and found a positive association between this measure of phonological similarity and participants' productive vocabulary sizes in their additional language. Using a similar measure of lexical similarity, Blom et al. (2019) extended these results to bilingual children aged three to 10, and reported a positive association between the lexical distance and comprehensive vocabulary size: bilinguals learning lexically close languages showed similar vocabulary sizes to those of monolinguals, but those learning the more distant languages showed lower vocabulary sizes.

These results suggest that linguistic distance between languages plays an important role during bilingual lexical acquisition, and that this effect operates within TEs. What mechanisms underlie this effect, and why they seem to operate differently across comprehension and production, or across the dominant and the non-dominant language, remain open issues. Floccia et al. (2018) pointed to parallel activation as a candidate mechanism behind this facilitation effect. The parallel activation principle suggests that bilinguals activate lexical representations in a language non-selective way: during the comprehension or production of a word in one language (e.g., *cat*, in English) its corresponding translation in the other language is activated too [e.g., *gato*, in Spanish. There is a vast body of evidence supporting language non-selective lexical access in both adults (e.g., Dell & O'Seaghdha, 1991; ???; Bobb, Von Holzen, Mayor, Mani, & Carreiras, 2020; Costa & Caramazza, 2000; Dijkstra, 2005; Hoshino & Kroll, 2008; Kroll, Gullifer, & Rossi, 2013; Singh, 2014; Thierry & Wu, 2007; Von Holzen, Fennell, & Mani, 2019; Yudes, Macizo, & Bajo, 2010; see Costa, Santesteban, & Caño, 2005 for review), and children

(e.g. Von Holzen & Mani, 2012; ???; ???; ???). A critical implication of parallel activation is that the word form (i.e., phonology, orthography, etc.) of both translation equivalents are available to each other, impacting the comprehension or production dynamic of any of them.

Floccia et al. (2018) argue that the degree of phonological similarity between TEs (*cognateness* from now on) of both languages should lead to increased parallel activation of both languages during language exposure, facilitating the acquisition of words in both languages, and leading participants learning two lexically close languages to show larger vocabulary sizes than those learning two lexically distant languages. This account, however, neglects the fact that for parallel activation to take place and play a facilitation role, lexical representations of both pairs of the TE must have been already established in the lexicon, and their corresponding phonological forms must have been encoded. This is not the case during lexical acquisition, where at least one of the members of the pair has not being acquired yet. A more lenient definition of parallel activation could adjust to the context of lexical acquisition more easily: when only one of the members of the TE has been acquired, its lexical representation is activated when, in the presence of its referent, a phonologically similar label is uttered. Consider the case of a Catalan-Spanish bilingual that has already learned the word *gat* (Catalan for *cat*), but not *gato* (in Spanish). It is possible that, when presented with *gato* in the presence of a cat, she maps this novel label to the familiar word *gat*, and creates a lexical representation for *gato* as a synonym for *gat*. The case of a non-cognate is differing. Consider now the case of a toddler that has already learnt *gos* (Catalan for *dog*), but not *perro* (Spanish). In this case, it would not be possible to map both labels via phonology, give their lack of phonological similarity.

Under this account, cognates (i.e., phonologically similar TEs) would be acquired earlier than non-cognates, but this effect would only play a role once one of the members of the TE has been acquired. The evidence supporting an earlier age of acquisition of cognates relative to non-cognates is sparse. Schelletter (2002) reported a longitudinal

case-study of a German/English child who produced cognate TEs earlier than non-cognate TE. Bosch and Ramon-Casas (2014), showed converging results from a sample of 48 monolingual and bilingual infants aged 24 months. The low number of words included in the analyses of both studies, and the lack of adjustment for lexical frequency, limits severely the strength of any conclusions concerning the difference in age of acquisition of cognates and non-cognates. On the other hand Floccia et al. (2018) worked with aggregated estimates of participants' vocabulary size, losing item-level information that could provide information about change in age of acquisition of TEs associated with their degree of phonological similarity. Therefore, their results can only be considered as compatible with cognateness playing a role on age of acquisition. In this study, we aimed at overcoming these pitfalls by using unaggregated data from individual responses, testing the role of cognateness explicitly and adjusting for lexical frequency.

It also follows from the fact that cognateness operates *within* TEs that the effect of cognateness can only play a role once one of the members of the TE has been acquired, and its phonological form is available during the acquisition of its translation. Words that belong to the language of most exposure (dominant language) are likely to be acquired earlier than words in the non-dominant language (Floccia et al., 2018). (???) tested Dutch-Frisian bilingual children aged 2.5 to four years in a comprehension task (PPVT-NL) that involved Frisian words with different degrees of phonological similarity with their Dutch TE. Infants with lower exposure to Frisian showed a better performance for words with high similarity, while no such benefit was found in children who were exposed mostly to Frisian. This suggests that infants that were mostly exposed to Dutch used the phonology of dutch words when processing Frisian words, as revealed by a better performance for cognates than for non-cognates. A study by (???) showed that the degree of phonological similarity between the performance of was better for cognates than for non-cognates, only for those children

Finally, previous studies have reported that the properties that describe the form of a

word play a stronger role in production than comprehension than in production. This is the case of the number of phonemes (Braginsky, Yurovsky, Marchman, & Frank, 2019), phonological neighbourhood density (Jones & Brandt, 2019). These results converge with Floccia et al. (2018)’s finding that the effect of the average cognateness between languages was larger in production than in comprehension¹. Accordingly, we expected the role of cognateness to be more central in production than in comprehension.

In summary, we investigated the effect of cognateness on the probability of acquisition of TEs, and predicted that (1) cognate TEs are acquired earlier than non-cognate TEs, that this effect will be larger in the dominant than in the non-dominant language, and larger in production than in comprehension.

Method

Participants

We collected data from 349 bilinguals (182 female), from the Metropolitan Area of Barcelona, between 28th October, 2019 and 09th January, 2021. All families gave informed consent before participating. This study was approved by the Comitè d’Ètica de la Investigació amb Medicaments (CEIm) from Hospital del Mar (Barcelona, Spain), code XXXXXXXXXX. We assessed toddlers’ language profile asking parents for an estimated proportion of exposure to each language. We excluded participants with >10% exposure to a third language. Fig. ?? illustrates the distribution of participants across language profiles and ages.

¹ Although Bosch and Ramon-Casas (2014) found that participants in their sample produced cognates earlier than non-cognates, no analyses were conducted on comprehensive data, and therefore they cannot be taken as supporting evidence for a larger effect of cognateness on production than in comprehension

Questionnaire

We implemented an on-line questionnaire using **formr** (Arslan, Walther, & Tata, 2020), divided in three forms: a (1) language questionnaire, a (2) demographic survey, and a (3) Catalan and a Spanish vocabulary checklists. Vocabulary checklists followed a similar structure as the Oxford Communicative Developmental Inventory (Hamilton et al., 2000) and consisted in two lists of words, one in Catalan and one in Spanish). Items in one language were translation equivalents of the items in the other (e.g., whenever *gos* [dog] was included in the Catalan inventory, the word *perro* was included in the Spanish inventory), roughly following a one-to-one mapping. When there were two acceptable translation equivalents for a given word, we included both (e.g., Catalan *acabar* [to finish] and Spanish *acabar* and *terminar*), or merged them into the same one (e.g., Spanish *mono* [monkey] and Catalan *mono/mico*). We included items from a diverse sample of 26 semantic/functional categories (see Appendix 1). We discarded the following categories: adverbs, auxiliary words, connectives, interjections and games and routines. The Catalan inventory contained 778 items (196 cognates, 582 non-cognates) and the Spanish inventory contained 781 (197 cognates, 584 non-cognates).

For each word in the vocabulary checklists, we asked parents to report whether their child was able to understand it, understand and say it, or did not understand or say it (marked by default). Participants filled a long or a short version of the questionnaire. Participants presented with the long version filled a list of 800 translation equivalents (800 items in Catalan and 800 items in Spanish), while participants presented with a short were randomly allocated into one of four list of items. Each list contained a different set of ~400 translation equivalents (~400 in Catalan, ~400 in Spanish). Semantic/functional categories with less than 16 items—thus resulting in less than four items after dividing it in four lists—were not divided in the short version of the questionnaire: all of their items were included in the four lists. Another subset of items that were part of the trial lists of some

experiments in the lab were also included in all versions. Table 2 in Appendix 1 shows the distribution of items across questionnaire versions. We excluded from the analysis multi-word items (e.g., *barrita de cereales* [cereal bar]) and items that included more than one word-form (e.g., *mono / mico*). Table ?? shows the classification of items in cognates and non-cognates and their frequency scores across the four lists of the inventories.

Data analysis

We gathered 89220 responses. Translation equivalents received an average of 189.83 ($SD = 67.27$, $Min = 75$, $Max = 312$), both languages summed together. We modelled the probability of children understanding or producing TEs separately using logistic regression model with a binomial distribution and a logit link function. (Agresti, 2010). We included several predictors of interest to adjust this probability to participants' and items' properties of interest. These predictors are:

- Age of participant in months (**age**, *Age*) calculated as the difference in days between participants' birth date and questionnaire completion divided by 30, chunked into bins of two months width.
- Degree of bilingualism of participant (**bilingualism**, *Bilingualism*) computed as the percentage of exposure to a second language (Spanish for participants exposed to >50% to Catalan and *vice versa*), chunked into six groups of participants, with degree of bilingualism 0%, 10%, 20%, 30%, 40% and 50%)
- Lexical frequency of item (**frequency**, *Frequency*) retrieved from SUBTLEX-CAT (Boada, Guasch, Haro, Demestre, & Ferré, 2019) for Catalan words and from SUBTLEX-ESP (Cuetos, Glez-Nosti, Barbón, & Brysbaert, 2011) for Spanish words, expressed as Zipf scores (Heuven, Mandera, Keuleers, & Brysbaert, 2014), and centred around the mean
- Language the item belongs to (**dominance**, *Dominance*) labelled as L1 if the item

belongs to the dominant language (e.g., a Catalan item is labelled as L1 for

Catalan-dominant participants and as L2 for Spanish-dominant participants)

- Cognateness of the item (**cognate**, *Cognate*), as rated by a trained Spanish-Catalan bilingual linguist.

We aggregated the data across participants. We summed the number of participants that *understood*, or *understood and produced* each TE for each combination of the **age**, **bilingualism** levels. Numerical variables (**age**, **bilingualism**, and **frequency**) were standardized, and categorical predictors were sum-coded. **dominance** was coded as L2 = -0.5 and L1 = +0.5, and **cognate** was coded as **Non-cognate** = -0.5 and **Cognate** = +0.5. In summary, each TE provided four data points per age bin: L1-Cognate, L1-Non-cognate, L2-Cognate, and L2-Non-cognate.

We adopted a Bayesian approach toward statistical inference. This approach allows to (1) incorporate previous domain knowledge into the inference process implementing the Bayes theorem, and (2) to quantify the uncertainty associated to the estimated parameters in our model (???). We first fit a base model that only adjusted for the age of participants and the lexical frequency of the items (??? @jones_do_2018). We then introduced our predictors of interest in the following order: (1) **dominance**, (2) **bilingualism**, (3) **dominance per bilingualism** interaction, (4) **cognate**, (5) **dominance per cognate** interaction, (6) **bilingualism per cognate** interaction, and (7) **dominance per bilingualism per cognate** interaction. Our extended model can be formalised as follows:

$$y_i \sim \text{Binomial}(n_i, p_i),$$

$$\text{logit}(p_i) = \alpha - \dots$$

$$\beta_1 \times \text{Age}_i + \dots$$

$$\beta_2 \times \text{Frequency}_i + \dots$$

$$\beta_3 \times \text{Dominance}_i + \dots$$

$$\beta_4 \times \text{Bilingualism}_i + \dots$$

$$\beta_5 \times \text{Cognate}_i + \dots$$

$$\beta_6 \times (\text{Dominance}_i \times \text{Bilingualism}_i) + \dots$$

$$\beta_7 \times (\text{Dominance}_i \times \text{Cognate}_i) + \dots$$

$$\beta_8 \times (\text{Bilingualism}_i \times \text{Cognate}_i) + \dots$$

$$\beta_9 \times (\text{Dominance}_i \times \text{Bilingualism}_i \times \text{Cognate}_i),$$

$$\text{logit}(p_i) = \log \frac{y_i}{n_i - y_i}$$

$$\alpha \sim \text{Normal}(0, 1.5)$$

$$\beta_{0-9} \sim \text{Normal}(0, 0.5)$$

Where:

- y_i is the sum of positive responses to *Understands* or *Understands and Says* for translation equivalent i
- n_i is the total number of responses for translation equivalent i
- α is the intercept, which corresponds to the probability of a positive response when all predictors equal 0
- $\beta_{1,\dots,9}$ are the coefficients of the predictors

We implemented this model using the R package **brms** (Bürkner, 2017) as **response**

~ age + frequency + dominance*bilingualism*cognate, running four MCMC

sampling chains with 1,000 iterations each, including 500 warm-up iterations per chain. Appendix 2 shows different diagnostics of chain convergence. We compared all models using leave-one-out cross-validation (LOO-CV) (Vehtari, Gelman, & Gabry, 2017)². This method computes, for a given model, the sum of the log scores of the posterior predictive distributions that result from removing one data-point at a time, providing an estimate of the model fit. The LOO estimates are adjusted for the number of parameters estimated in the model, therefore accounting for overfitting. We compared the expected log posterior density of each each model to test which one fitted the data the best.

We then explored the posterior distribution of each parameter in the model that fitted the data the best computing the 95% credible intervals and testing whether this interval excluded 0. Credible intervals (CrI) indicate the range of values we are 95% certain contain the true value of the parameter, given our prior and observed data. We performed follow-up tests on interactions whose 95% credible interval excluded 0 by comparing the 95% credible interval of their estimated marginal means using the `emmeans` package (Lenth, 2020). Data processing and visualisation was done in R (R Core Team, 2020) using the `tidyverse` family of packages (Wickham et al., 2019), and posterior samples were extracted using the `tidybayes` R package (Kay, 2020).

Results

Comprehension

The extended model, including the three-way interaction `dominance1:bilingualism;cognate1` fitted comprehension data the best: it showed the smallest absolute expected log predictive density. This estimate is several times larger than its standard error indicating high predictive accuracy. Table 2 summarises the LOO scores

² Due to the high computational cost associated with a large dataset, we used a sub-sampling approach for performing Bayesian LOO with 1,000 samples (Magnusson, Andersen, Jonasson, & Vehtari, 2019)

282 of the fitted models.

Model	ELPD ^a	SE ^b	p ^c	Comparison	
				<i>LOOdiff</i>	<i>SEdiff</i>
Production					
~1 + age + frequency	−73,655.36	277.24	5.93	4577.29336	380.0306
... + dominance	−71,344.56	271.05	7.19	2266.48651	375.5426
... + bilingualism	−70,706.33	267.57	9.23	1628.26183	373.0344
... + dominance:bilingualism	−69,246.49	261.81	11.57	168.42565	368.9302
... + cognate	−69,242.57	261.70	10.65	164.49868	368.8513
... + dominance:cognate	−69,143.92	260.65	13.29	65.84950	368.1056
... + bilingualism:cognate	−69,129.05	260.58	14.47	50.98597	368.0542
... + dominance:bilingualism:cognate	−69,078.07	259.93	15.72	0.00000	0.0000
Comprehension					
~1 + age + frequency	−73,653.35	277.26	5.03	4577.41755	380.0019
... + dominance	−71,348.67	271.01	6.44	2272.73124	375.4690
... + bilingualism	−70,702.30	267.45	6.57	1626.36697	372.9028
... + dominance:bilingualism	−69,245.99	261.71	8.87	170.05482	368.8094
... + cognate	−69,240.75	261.61	9.43	164.81145	368.7373
... + dominance:cognate	−69,142.47	260.63	9.67	66.53140	368.0479
... + bilingualism:cognate	−69,134.13	260.64	15.77	58.19037	368.0546
... + dominance:bilingualism:cognate	−69,075.94	259.86	17.30	0.00000	0.0000

^aELPD: theoretical expected log pointwise predictive density for a new dataset, as estimated using Bayesian LOO-CV

^bSE: standard error of the ELPD

^cp: effective number of parameters

We report the mean of each coefficient's posterior distribution along with standard errors and 95% CrI in Table 3. For interpretability, we report the derivative of the logistic function³ for each coefficient, which indicates the maximum difference in probability corresponding to one unit different in the input variable (??).

Coefficient	Comprehension			Production		
	<i>Mean</i>	<i>SE</i> ¹	<i>95% CrI</i> ²	<i>Mean</i>	<i>SE</i> ¹	<i>95% CrI</i> ²
Intercept	0.48	0.50	0.48–0.48	0.14	0.50	0.14–0.14
Age	0.76	0.50	0.75–0.76	0.81	0.50	0.81–0.81
Frequency	0.61	0.50	0.61–0.62	0.61	0.50	0.61–0.61
Dominance	0.68	0.50	0.68–0.69	0.71	0.50	0.71–0.72
Bilingualism	0.55	0.50	0.55–0.56	0.54	0.50	0.53–0.54
Cognate	0.51	0.50	0.50–0.51	0.55	0.50	0.54–0.55
Dominance * Bilingualism	0.34	0.50	0.34–0.35	0.37	0.50	0.37–0.38
Dominance * Cognate	0.42	0.51	0.41–0.43	0.39	0.51	0.38–0.41
Bilingualism * Cognate	0.48	0.50	0.48–0.49	0.49	0.50	0.48–0.50
Dominance * Bilingualism * Cognate	0.56	0.51	0.55–0.57	0.56	0.51	0.54–0.57

¹SE: Standard Error of the mean

²95% CrI: 95% credible interval

The intercept (47.98%) indicates probability of participants understanding TEs at mean age (25.48 months), mean frequency (2.93 Zipf score), and mean L2 exposure (24.22%), across the levels of the **dominance** and **cognate** factors. An increase in one standard deviation increment in age (1 *SD* = 7.30 months) is associated with a 28.20% increment in comprehension probability. A one standard deviation increase in lexical frequency (1 *SD* = 1.41 Zipf scores) is associated with a 11.49% increment in

³ The derivative is calculated as $\hat{\beta}_j/4$, where $\hat{\beta}_j$ is the estimated mean of the posterior distribution of coefficient j . This value corresponds to the slope of the logistic curve at the midpoint, where it is steepest.

comprehension probability. Items in the dominant language were 19.23% more likely to be understood than items in the non-dominant language. A one standard deviation increase in exposure to the non-dominant language ($1\ SD = \text{printnum(sd(responses\$bilingualism_scaled))\%}$) is associated with a 5.48 increment in comprehension probability. Cognates were 0.62% more like to be understood than non-cognates. The 95% credible interval of the three-way interaction (`dominance:bilingualism:cognate`) excluded 0, indicating that that the cognate status of the items changed the effect of dominance differently depending on the amount of exposure to the second language. Figure 1 shows the estimated marginal mean posterior probability of comprehension and production across ages, and the different levels of `dominance`, `bilingualism`, and `cognate`.

Children exposed to their second language more than -1 SD away from the mean (7.33%) were more likely to understand cognates than non-cognates in their non-dominance language ($M_{Cognate} - M_{Non-cognate} = 9.61\%$, 95% CrI = [8.33-10.87]), but not in their dominant language ($M_{Cognate} - M_{Non-cognate} = -5.11\%$, 95% CrI = [-6.14-4.07]). This difference was smaller for participants with mean exposure to the non-dominant language (24.22%): in the non-dominant language cognates were slight more likely to be understood than non-cognates ($M_{Cognate} - M_{Non-cognate} = 4.83\%$, 95% CrI = [3.96-5.69]), but not in the dominant language ($M_{Cognate} - M_{Non-cognate} = -3.58\%$, 95% CrI = [-4.36-2.81]). Finally this effect was negligible for participants exposed to their second language more than -1 SD away from the mean (41.11%). Cognates and non-cognates were equally likely to be understood in both the non-dominant ($M_{Cognate} - M_{Non-cognate} = 0.05\%$, 95% CrI = [-1.18-1.19]), and the dominant language ($M_{Cognate} - M_{Non-cognate} = -2.05\%$, 95% CrI = [-3.20-0.91]).

The difference in probability of comprehension between the dominant and the non-dominant languages differed between cognates and non-cognates and different levels of exposure to the non-dominant language. For children exposed. For children exposed to

their non-dominant language less than 1 SD from the mean, difference in probability between the dominant and the non-dominant language was smaller for cognates ($M_{Dominant} - M_{Non-dominant} = 27.97\%$, 95% CrI = [27.06-28.88]) than for non-cognates ($M_{Dominant} - M_{Non-dominant} = 42.68\%$, 95% CrI = [41.31-44.03]). For children with mean exposure to the non-dominant language, this effect was smaller: the difference between the comprehension probability of items in L1 and L2 was slightly smaller for cognates ($M_{Dominant} - M_{Non-dominant} = 15.03\%$, 95% CrI = [14.39-15.68]) than for non-cognates ($M_{Dominant} - M_{Non-dominant} = 23.44\%$, 95% CrI = [22.48-24.37]). Finally this effect was negligible for children exposed more than 1 SD than the mean to their non-dominant language: the difference in comprehension probability between items in the dominant and the non-dominant language was roughly similar for cognates ($M_{Dominant} - M_{Non-dominant} = 2.09\%$, 95% CrI = [1.19-3.03]) and non-cognates ($M_{Dominant} - M_{Non-dominant} = 4.19\%$, 95% CrI = [2.81-5.58]).

Discussion

References

- Agresti, A. (2010). *Analysis of ordinal categorical data* (Vol. 656). John Wiley & Sons.
- Arslan, R. C., Walther, M. P., & Tata, C. S. (2020). Formr: A study framework allowing for automated feedback generation and complex longitudinal experience-sampling studies using R. *Behavior Research Methods*, 52(1), 376–387. <https://doi.org/10.3758/s13428-019-01236-y>
- Ben-Zeev, S. (1977). The Influence of Bilingualism on Cognitive Strategy and Cognitive Development. *Child Development*, 48(3), 1009–1018. <https://doi.org/10.2307/1128353>
- Bialystok, E., Luk, G., Peets, K. F., & Yang, S. (2010). Receptive vocabulary

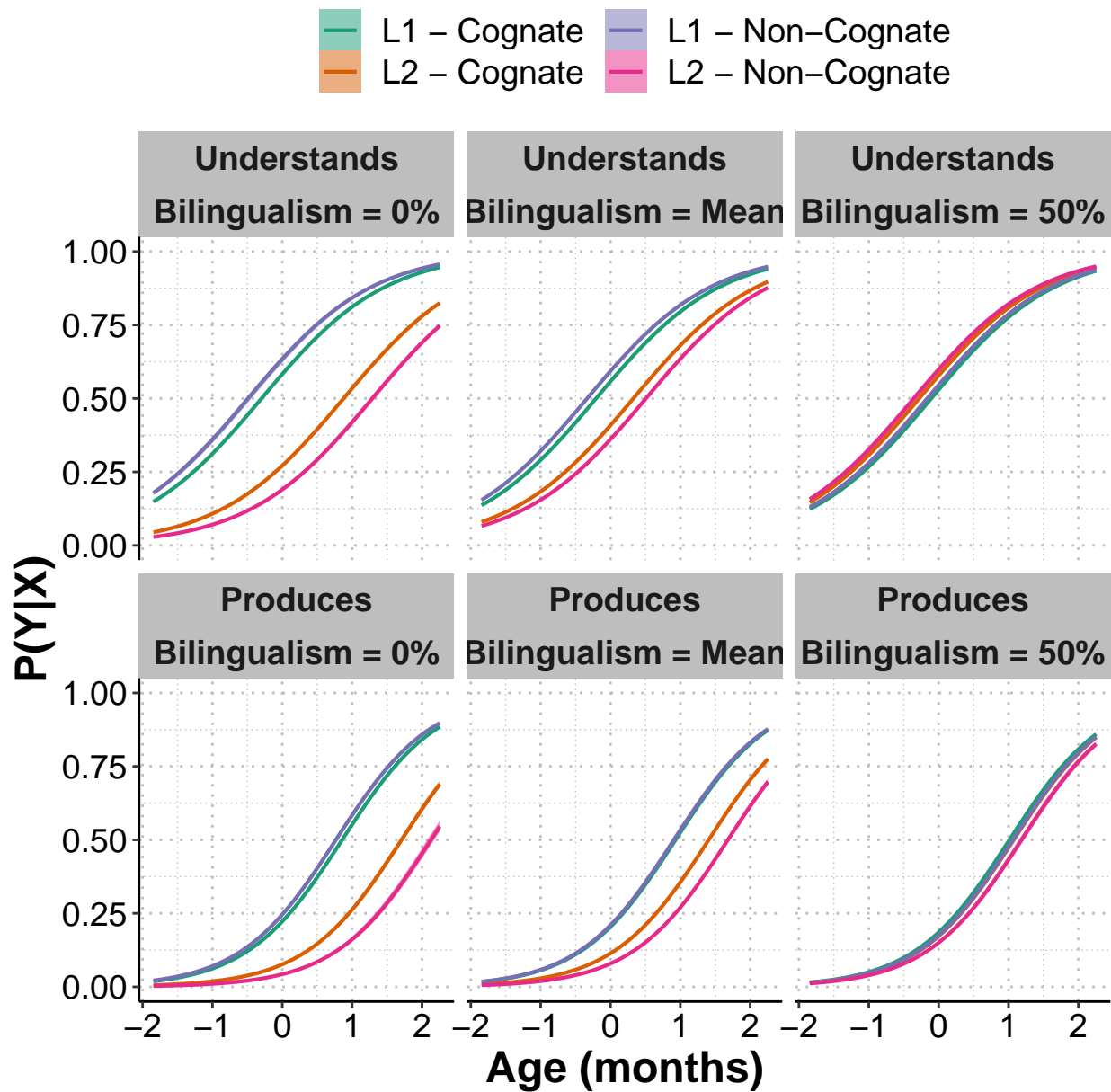


Figure 1. Mean posterior probability and 95% CrI of responses, conditional to Dominance, Bilingualism, and Cognate.

differences in monolingual and bilingual children. *Bilingualism: Language and Cognition*, 13(4), 525–531. <https://doi.org/10.1017/S1366728909990423>

Blom, E., Boerma, T., Bosma, E., Cornips, L., Heuij, K. van den, & Timmermeister, M. (2019). Cross-language distance influences receptive vocabulary outcomes of bilingual children. *First Language*, 0142723719892794. <https://doi.org/10.1177/0142723719892794>

Boada, R., Guasch, M., Haro, J., Demestre, J., & Ferré, P. (2019). SUBTLEX-CAT: Subtitle word frequencies and contextual diversity for Catalan. *Behavior Research Methods*. <https://doi.org/10.3758/s13428-019-01233-1>

Bobb, S. C., Von Holzen, K., Mayor, J., Mani, N., & Carreiras, M. (2020). Co-activation of the L2 during L1 auditory processing: An ERP cross-modal priming study. *Brain and Language*, 203, 104739. <https://doi.org/10.1016/j.bandl.2019.104739>

Bosch, L., & Ramon-Casas, M. (2014). First translation equivalents in bilingual toddlers' expressive vocabulary: Does form similarity matter? *International Journal of Behavioral Development*, 38(4), 317–322. <https://doi.org/10.1177/0165025414532559>

Braginsky, M., Yurovsky, D., Marchman, V. A., & Frank, M. C. (2019). Consistency and Variability in Children's Word Learning Across Languages. *Open Mind*, 3, 52–67. https://doi.org/10.1162/opmi_a_00026

Bürkner, P.-C. (2017). Brms: An R Package for Bayesian Multilevel Models Using Stan. *Journal of Statistical Software*, 80(1), 1–28. <https://doi.org/10.18637/jss.v080.i01>

Core, C., Hoff, E., Rumiche, R., & Señor, M. (2013). Total and Conceptual Vocabulary in Spanish–English Bilinguals From 22 to 30 Months: Implications for Assessment. *Journal of Speech, Language, and Hearing Research*, 56(5),

1637–1649. [https://doi.org/10.1044/1092-4388\(2013/11-0044\)](https://doi.org/10.1044/1092-4388(2013/11-0044))

Costa, A., & Caramazza, A. (2000). Lexical access in speech production: The bilingual case. *Psicológica*, 21(2), 403–437.

Costa, A., Santesteban, M., & Caño, A. (2005). On the facilitatory effects of cognate words in bilingual speech production. *Brain and Language*, 94(1), 94–103.

Cuetos, F., Glez-Nosti, M., Barbón, A., & Brysbaert, M. (2011). SUBTLEX-ESP: Spanish word frequencies based on film subtitles. *Psicológica*, 32(2), 133–143.

Dell, G. S., & O'Seaghdha, P. G. (1991). Mediated and convergent lexical priming in language production: A comment on levelt et al (1991).

Dijkstra, T. (2005). Bilingual Visual Word Recognition and Lexical Access. In *Handbook of bilingualism: Psycholinguistic approaches* (pp. 179–201). New York, NY, US: Oxford University Press.

Fernandez, M. C., Pearson, B. Z., Umbel, V. M., Oiler, D. K., & Molinet-Molina, M. (1992). Bilingual Receptive Vocabulary in Hispanic Preschool Children. *Hispanic Journal of Behavioral Sciences*, 14(2), 268–276.
<https://doi.org/10.1177/07399863920142006>

Floccia, C., Sambrook, T. D., Luche, C. D., Kwok, R., Goslin, J., White, L., ... Plunkett, K. (2018). I: Introduction. *Monographs of the Society for Research in Child Development*, 83(1), 7–29. <https://doi.org/10.1111/mono.12348>

Hamilton, A., Plunkett, K., & Schafer, G. (2000). Infant vocabulary development assessed with a british communicative development inventory. *Journal of Child Language*, 27(3), 689–705.

Heuven, W. J. B. van, Mandera, P., Keuleers, E., & Brysbaert, M. (2014). Subtlex-UK: A New and Improved Word Frequency Database for British English: *Quarterly Journal of Experimental Psychology*. Retrieved from

<https://journals.sagepub.com/doi/10.1080/17470218.2013.850521>

Hills, T. T., Maouene, M., Maouene, J., Sheya, A., & Smith, L. (2009).

Longitudinal Analysis of Early Semantic Networks Preferential Attachment or
Preferential Acquisition? *Psychological Science*, 20(6), 729–739.

<https://doi.org/10.1111/j.1467-9280.2009.02365.x>

Hoff, E., Core, C., Place, S., Rumiche, R., Señor, M., & Parra, M. (2012). Dual
language exposure and early bilingual development*. *Journal of Child Language*,
39(1), 1–27. <https://doi.org/10.1017/S0305000910000759>

Hoff, E., Rumiche, R., Burridge, A., Ribot, K. M., & Welsh, S. N. (2014).

Expressive vocabulary development in children from bilingual and monolingual
homes: A longitudinal study from two to four years. *Early Childhood Research*
Quarterly, 29(4), 433–444. <https://doi.org/10.1016/j.ecresq.2014.04.012>

Hoshino, N., & Kroll, J. F. (2008). Cognate effects in picture naming: Does
cross-language activation survive a change of script? *Cognition*, 106(1), 501–511.
<https://doi.org/10.1016/j.cognition.2007.02.001>

Houwer, A. D., Bornstein, M. H., & Putnick, D. L. (2014). A bilingual–monolingual
comparison of young children’s vocabulary size: Evidence from comprehension
and production. *Applied Psycholinguistics*, 35(6), 1189–1211.
<https://doi.org/10.1017/S0142716412000744>

Jones, S. D., & Brandt, S. (2019). Do children really acquire dense neighbourhoods?
Journal of Child Language, 46(6), 1260–1273.
<https://doi.org/10.1017/S0305000919000473>

Kay, M. (2020). *tidybayes: Tidy data and geoms for Bayesian models*.
<https://doi.org/10.5281/zenodo.1308151>

Kroll, J. F., Gullifer, J. W., & Rossi, E. (2013). The Multilingual Lexicon: The

Cognitive and Neural Basis of Lexical Comprehension and Production in Two or More Languages. *Annual Review of Applied Linguistics*, 33, 102–127.

<https://doi.org/10.1017/S0267190513000111>

Lenth, R. V. (2020). *Emmeans: Estimated marginal means, aka least-squares means*. Retrieved from <https://CRAN.R-project.org/package=emmeans>

Magnusson, M., Andersen, M., Jonasson, J., & Vehtari, A. (2019). Bayesian leave-one-out cross-validation for large data. In K. Chaudhuri & R. Salakhutdinov (Eds.), *Proceedings of the 36th international conference on machine learning* (Vol. 97, pp. 4244–4253). Long Beach, California, USA: PMLR. Retrieved from <http://proceedings.mlr.press/v97/magnusson19a.html>

Pearson, B. Z., & Fernández, S. C. (1994). Patterns of Interaction in the Lexical Growth in Two Languages of Bilingual Infants and Toddlers. *Language Learning*, 44(4), 617–653. <https://doi.org/10.1111/j.1467-1770.1994.tb00633.x>

Pearson, B. Z., Fernández, S. C., & Oller, D. K. (1993). Lexical Development in Bilingual Infants and Toddlers: Comparison to Monolingual Norms. *Language Learning*, 43(1), 93–120. <https://doi.org/10.1111/j.1467-1770.1993.tb00174.x>

R Core Team. (2020). *R: A language and environment for statistical computing*. Vienna, Austria: R Foundation for Statistical Computing. Retrieved from <https://www.R-project.org/>

Schelleter, C. (2002). The effect of form similarity on bilingual children's lexical development. *Bilingualism: Language and Cognition*, 5(2), 93–107. <https://doi.org/10.1017/S1366728902000214>

Singh, L. (2014). One world, two languages: Cross-language semantic priming in bilingual toddlers. *Child Development*, 85(2), 755–766.

Thierry, G., & Wu, Y. J. (2007). Brain potentials reveal unconscious translation

during foreign-language comprehension. *Proceedings of the National Academy of Sciences*, 104(30), 12530–12535. <https://doi.org/10.1073/pnas.0609927104>

Vehtari, A., Gelman, A., & Gabry, J. (2017). Practical Bayesian model evaluation using leave-one-out cross-validation and WAIC. *Statistics and Computing*, 27(5), 1413–1432. <https://doi.org/10.1007/s11222-016-9696-4>

Von Holzen, K., Fennell, C. T., & Mani, N. (2019). The impact of cross-language phonological overlap on bilingual and monolingual toddlers' word recognition. *Bilingualism: Language and Cognition*, 22(3), 476–499. <https://doi.org/10.1017/S1366728918000597>

Von Holzen, K., & Mani, N. (2012). Language nonselective lexical access in bilingual toddlers. *Journal of Experimental Child Psychology*, 113(4), 569–586. <https://doi.org/10.1016/j.jecp.2012.08.001>

Wickham, H., Averick, M., Bryan, J., Chang, W., McGowan, L. D., François, R., ... Yutani, H. (2019). Welcome to the `tidyverse`. *Journal of Open Source Software*, 4(43), 1686. <https://doi.org/10.21105/joss.01686>

Yudes, C., Macizo, P., & Bajo, T. (2010). Cognate effects in bilingual language comprehension tasks. *NeuroReport*, 21(7), 507–512. <https://doi.org/10.1097/WNR.0b013e328338b9e1>